# A STRUCTURE FOR CONSCIOUS THOUGHT:

# THE ARTIFICIAL INTELLIGENCE PERSPECTIVE

Peter W. Wallace
Mind-Link Research Institute
898 14th Street
San Francisco CA 94114
(415) 861-7671

**About the Author:** Peter Wallace holds the B.S. in Cognitive Science from the Massachusetts Institute of Technology and the M.S. in Engineering-Economic Systems from Stanford University. He has worked extensively in the field of Artificial Intelligence and is currently with the Mind-Link Research Institute in San Francisco.

Each of the sciences has its own perspective on the study of consciousness, and Artificial Intelligence is no exception. Although the terminology and fundamental concerns of each of these fields will presumably differ, the multiple viewpoints surely complement one another and contribute to a robust understanding. Using the analogy of the study of video cameras, we can say that if Biology tells us the reasons behind the development of video cameras, and Physics tells us how people's behavior differs when they are being filmed, then the task falls to Artificial Intelligence to explain the camera's internal workings.

Several major figures in the Artificial Intelligence field have written at length on the topic of consciousness, including Marvin Minsky of M.I.T. and Douglas Hofstadter of the University of Indiana. Minsky, in The Society of Mind (1986) and elsewhere, focuses on elaborating the multiple functional agents that, he feels, compose our mental life. He summarizes his views on the mind-body problem as "Minds are simply what brains do." Hofstadter (1980) has written extensively on the strange phenomena of self-perception, or as he refers to them "Strange Loops," involved in our being conscious of our own thinking, or as in Godel's Incompleteness Theorem, with any formal system speaking about itself.

In addition to these efforts, we should mention one of the computational architectures used for the development of expert systems, the BLACKBOARD architecture (Englemore, 1988). Englemore describes this approach, which utilizes a centralized "blackboard" that is perceived and commented on by numerous mental agents, as being the most flexible architecture yet found for expert system development. Originally devised for the (presumably) non-conscious process of speech recognition, it also provides a suggestive metaphor for consciousness, which we will claim is our most flexible mode of thinking.

The Artificial Intelligence approach to examining consciousness is the same functionalist approach used throughout the computer sciences - divide the phenomenon to be studied into distinct functional subunits and trace the communication between them. In what follows we will attempt a first cut at this type of analysis.

Another basic tenet of explanation is that a property to be explained should be built up from elements that do not themselves exhibit that property (Hofstadter, 1985). For instance, an explanation of why something appears green (the property greenness) should not refer to green molecules, but instead to how the molecules radiate light at a certain wavelength. Similarly, our account of consciousness will involve the interaction of numerous mental "agents" and these agents should specifically not be attributed with consciousness. We need to avoid the well-known fallacy of postulating a homonculus - a "little man in the head" who looks out at the world and decides what to do. Our account, therefore, will hypothesize only primitive, non-conscious entities, perhaps no more sophisticated than today's expert systems, whose pattern-matching capabilities are relatively transparent and comprehensible without reference to consciousness. We will propose that just as the property green can arise from a collection of colorless molecules, the phenomenon of consciousness may arise from the interaction of non-conscious functional units.

In what follows, we will describe three case studies, each of which illuminates some aspect of consciousness. The first of these, reading to ourselves, is an everyday experience. The other two, functioning at an expert level during competition, and being caught in an infinite loop, are less commonplace and yield more privileged insights. We will then attempt to synthesize what we have learned, by proposing a particular structure for consciousness that exhibits the required properties. In particular, we will derive a structure that accounts for the peculiar process whereby it is possible for you to hear yourself think.

**Case Study 1: Reading to Yourself**

An easily accessible and replicable activity that summons forth the "little voice in your head" is to read to yourself. While you read to yourself, the voice in your head is speaking thoughts that originate outside of yourself. If you are attentive, however, it is also possible to observe it voicing your personal reactions to what you read.

It is useful to distinguish several possible reactions one can have while reading. The first is typical for reading material that we do not find inherently interesting, but are required to read anyway - an accounting textbook might be a good example. This type of response I would characterize as "Yes, I understand, go on." The second is typical of material that we may find quite interesting, but do not have much to say about, such as a good spy thriller. Here our response can be stereotyped as "What's next? What's next?". In both of these cases I would claim that while we are aware of the material, we have few conscious thoughts about it.

In a third case, such as an interesting paper in a field we know something about, we do have conscious reactions, which are the type of comments that we might put in the margin. These I would characterize as "Let me comment here that ...". The final type, which is of particular interest for this paper, involves statements that provoke such a reaction in us that we cannot comprehend our reaction all at once. What reaction, for instance, does one have to reading that when they were kids, Ronald Reagan and his adopted baby brother, Timothy Leary, struck a deal that they would someday alter the American consciousness. I would suggest that to the extent that the statement is taken seriously, it will provoke multiple reactions involving several previously distinct topics - this is news about Reagan and news about Leary simultaneously, as well as about other topics. The thoughts cannot be all made conscious at once, but if one pauses a moment, and perhaps rereads the statement, it is possible to step through each of them. Although on occasion some of the thoughts can even lead to "cascades" of further thoughts, the cascades typically do not go very far, and in a short time we are free to move on to the rest of the thoughts. This type of reaction I characterize as "Not so fast - let me think about that a moment." This last is a vivid example of the subjective experience commonly referred to as having one's head crowded with thoughts. In more technical terms we would say that the thoughts interfere with each other, and would deduce that they are competing for some shared resource. How it is possible to generate more thoughts that we can handle will be an issue for any proposed model to resolve.

**Case Study 2:** **Sitting for the Math Exams**

Two incidents, both involving math exams, provide an in-depth illustration of the differences between the modes of thought we can operate in, the conscious and the non-conscious. A little of this difference can be appreciated by comparing the subjective experiences of 1) verifying that "six times eight is forty-eight" and 2) verifying that "three times two times two times four is forty-eight". For most people, the first answer comes into their head without conscious thought, whereas the second must be reached through a series of intermediate conscious steps, something like "three times two is six, times two is twelve, times four is forty-eight." This example is inadequate, though, as it might be taken to be just the difference between remembering something and figuring it out. The incidents below will illustrate the principle in greater depth.

An incident my senior-year in high school first focused my attention on the difference between conscious and non-conscious thought. I was involved in a national competition in mathematics for high-school students, and I had reached an expert level of ability where I stood a good chance to get the perfect score that would guarantee first place in the country. When I took the actual test, the process of solving the problems was subjectively similar to what it had been in the preliminaries, a few obvious steps and then the answer. With five minutes to go and one question remaining, though, something went wrong, and the steps did not come to me automatically. Frantically, I racked my brain for the answer, and with a minute to go the insight came to me that solved the problem. Still, mixed with the relief and excitement was an amazement at the immeasureable gulf between the two modes of thought that had been exhibited during the test, and it was this that first made me ask who does the thinking in there.

Five years later, this issue of who does the thinking was brought into sharper focus during the entrance examinations for graduate school. As I started on the mathematics section I found that no sooner would I finish reading a question (to myself) than I would hear a voice say the answer (internally) as clearly as if someone were whispering it in my ear. I suspect that this experience is probably not so different from what people are referring to when they say their Muse speaks to them, but for me to use that terminology here would be premature. After several such questions I became concerned - how was I to know that these "instant answers" were correct? After all, it had been a long time since I had done problems of this sort (primarily pre-college level mathematics).

If I wanted to maximize my chances of a perfect score, perhaps I had better think about what I was doing. For the next fifteen minutes, I attempted to be as conscious as possible of the intermediate steps I was taking, to take steps sufficiently small that I could be confident that they were within my competence, and to only actually take the steps after I had "verified" them. But this "verification" process turned out to simply consist of holding the reasoning step in my consciousness and seeing if I objected to it - that is, seeing if I could spot anything wrong with it. I was anticipating that, for instance, at some point I would be about to take a step like "and 7 times 8 is 54..." and realize, "oops, that was a close call, 7 times 8 is not 54, it's 56." Nothing of this sort happened at the time, and my current belief is that, except in special circumstances, there is no source for a "second opinion". If you think that 7 times 8 is 54 the first time, it will probably still seem right to you on reexamination.

The process of attempting to consciously verify the reasoning was, however, debilitatingly slow, so that with half the time gone and too few questions answered, I realized that my best bet at that point was to return to "instant answers" mode and hope for the best. As it happened, the instant answers were error-free. In doing a follow-up investigation, however, of the issues raised by this incident, I found that the "instant answers" were susceptible to some faulty reasoning steps, which conscious examination could catch, such as "The square-root of 9 is 3" instead of "The square-roots of 9 are +3 and -3". In this instance, the "instant answers" would assume that numbers only had positive square roots (unless there were strong contextual reminders of the negative square root as well). Yet, when the steps were reviewed consciously, the error was obvious. Fortunately, this was not an issue on the graduate record exam, but it does illustrate a unique property of consciousness - that holding something in consciousness can occasionally make it visible to extra knowledge sources from which it would normally be insulated. It would appear that at least in some cases consciousness serves as some kind of interface between knowledge sources that would not otherwise be able to communicate.

### Case Study 3: Caught in an Infinite Loop

Our final case study involves a case where several out-of-the-ordinary circumstances led to an unusual glimpse into the workings of subjective mental life. These circumstances, which included some erroneous beliefs that I held at the time, led to an episode where I literally found my mind trapped in the pathological state known to computer programmers as an "infinite loop".

The incident occurred in an environment widely reported to bring about non-ordinary experiences of consciousness - the sensory-deprivation flotation tank. The use of the tank was pioneered by John C. Lilly, M.D. in the 1950's, and has been widely used since. The physicist Richard Feynman discusses his flotation tank experiences at length in his book Adventures of a Curious Character, where he describes the unusual flavor that one's mental processes take on when deprived of their usual input.

The tank itself consists of a chamber filled with water just warm enough (93°) to remove any sensation of hot or cold, and sufficiently salinated that the user will be positively buoyant. While floating in the chamber, one is enveloped by total darkness and other silence, leaving little for one to be conscious of, except one's own thoughts. It was while floating in such an environment that I became trapped in an infinite loop.

At the time I had been working on a lot of computer programming, and had recently witnessed several instances of programs going into infinite loops. The use of repetitive loops is fairly commonplace in procedural programming languages, specifically to bring about the repeated execution of a series of instructions. It is expected that after a sufficient number of repetitions, a pre-specified test condition will be met, and the program will move onto some other task. One may think of each repetition, then, as making progress toward satisfying the test condition, but in the occasional pathological program, the pre-specified condition can never be met, and the program will repeat the instructions forever, each execution merely setting the stage for the next, without any hope of moving on to another task. Getting trapped in such a loop renders the host computer useless until it is unplugged or otherwise reset.

Having seen computers get caught in this process, and convinced that the analogy between human and computer thought was a quite literal one, I began to speculate, as I floated in the tank, whether humans might not be susceptible to similar pathologies. Since thoughts seem to trigger other thoughts, I reasoned, a thought that can trigger an identical thought could put a person into an infinite loop. Having once had the thought, the person would be doomed to having that same thought over and over again, just like their computer counterpart. Yes, I concluded, a person can get trapped in an infinite loop. Foolishly, I posed myself the question: what would this pathological thought be?

After a moment's reflection the answer came to me: "THIS THOUGHT WILL REPEAT ITSELF FOREVER". My first reaction was, of course, to verify this answer, and I concluded, "Yes, this thought will repeat itself forever," at which point I realized "Oh no, I've thought the demon-thought, and now it really will repeat itself forever, in my head!" Minute by minute, with mounting panic, various verbalizations of this same idea raced through my mind. Strangely, no two sentences were exactly the same, but each contained the same underlying idea: I was caught in an infinite loop and I was going to keep thinking this thought forever. Each thought set the stage for its successor, with each iteration moving fluidly from being something I thought to being something I thought about.

A possibility I hadn't foreseen saved the day - I got distracted. Without even having been aware of the shift of attention, I found myself contemplating an appointment scheduled for the next day. When it dawned on me that, miraculously, I had escaped the loop, I realized that my initial reasoning must have been fallacious, that the mind is probably not capable of thinking the same thought literally forever, that distraction or fatigue would surely always save the day. For this reason, I need never fear stumbling on that thought again, because it is no longer its own stimulus, but is, rather, a stimulus for the comment, "No, that's not true, that thought might repeat itself for quite a while, but we know it won't repeat forever." This new thought, obviously, might also be capable of causing self-repetition in the correct circumstances, but without the driving panic (that it might repeat forever), I find it doubtful that any one thought could dominate one's attention for all that long.

This once-in-a-lifetime incident provides a wealth of data and has some fascinating implications for any proposed model of conscious thought. True, the mind was after all sufficiently different from a computer that it was able to exit the loop without outside intervention. Still, after ten years working in the field of Artificial Intelligence, and keeping in mind the wealth of differences that have been uncovered between the operations of brains and the computers of today, I am frankly amazed that the mind has such a structure that it can in fact get caught in loops at all.

## IMPLICATIONS

Let us summarize what we have learned from our case studies.

From the experience of having our minds "crowded" with thoughts, as happens sometimes when reading or when stimulated by a particularly provocative remark, we see that there is definitely more than one source of ideas and comments in our heads, that these multiple agents can generate ideas in parallel, and that these ideas cannot all be aired consciously at once.

In the case study of the math exams we saw illustrated in detail the difference between conscious and non-conscious thought. It appears that the same mental steps can be traced through in either of the two modes, but that certain objections to the reasoning can only be made when the processing is conscious. Reasoning at the conscious level seems to allow additional interaction of some parts of our knowledge with other parts of our knowledge. We can thus characterize conscious reasoning as being in some sense "public", and subject to scrutiny. By contrast, reasoning steps taken outside of consciousness can be considered in some sense "private" (and non-correctable).

The episode of the infinite loop illustrates poignantly that in the act of consciously thinking a thought that thought becomes externalized to where it can itself be the object of further thought. Like the people who half-jokingly say, "How do I know what I think until I've heard what I've said?" we seem to listen to our internal voice to hear what it is that we think, and once we've heard it, our statement is an object for scrutiny like anyone else's statements. That a series of (identical) thoughts could provoke a series of identical reactions demonstrates a certain mind-less, robotic unselfconsciousness on the part of whatever is doing the thinking. Nothing allows the thought-source to notice that its last three comments were the same. The source seems not to realize that it is the source of the comments, nor that it is the cause of the infinite loop, nor that the alarm has already been sounded. Like the watchdog barking at its own echo, it persists in sounding its warning.

How, then, to synthesize these various insights? What structure can we hypothesize for consciousness that will reflect these considerations? Together, I believe, these constraints point to an architecture that is without a real parallel in the world we are familiar with. Still, the elements are familiar to us, and we can, therefore, construct a metaphorical model of the architecture we propose.

### THE SINGLE-SPOKESMAN MODEL

The considerations we have raised point to a "single spokesman - many commentators" structure for conscious thought. The appropriate image is of a huge auditorium filled with "commentator-agents", and a single "spokesman" on the stage at the front. The spokesman originates no thoughts whatsoever, but instead reads aloud notes sent to him from members of the audience. I "hear myself think" exactly when this spokesman reads my thoughts aloud in my head. The members of the audience, in turn, listen to the spokesman, and when appropriate, they comment on his remarks. (It is assumed that at most times few, if any, of the audience have a comment to make.) Now this is the circular part, which makes infinite loops possible: the process of commenting consists simply of writing a note and sending it to the front to read aloud. The commentators have no voices of their own, and must always express themselves through the single, shared spokesman. It is through this process that we become conscious of our thoughts. It is this spokesman that you hear when you hear yourself think.

In this view, being conscious of your own thoughts is very similar to being conscious of verbal input from the outside world. In both cases, "you" hear a voice speaking, and the full range of your mental resources is brought to bear on what you are hearing. In terms of our model, we would say that the statement is made public to the members of the audience. The difference is that in the latter case it is someone else speaking, while in the first it is your Self. The case of reading to yourself is a hybrid of these two extremes: the thoughts are someone else's, but it is your spokesman that is articulating them. In none of these cases is their a specific one of the commentators that can be singled out as "the Real You" doing the listening: You-the-Conglomerate hears the voice exactly when all the little commentators hear it. However, if it makes sense to speak of You hearing someone else speak to you, it makes just as much sense to speak of You hearing the internal spokesman.

Probably the strangest feature of this model is the relationship it suggests between the perceiver and the perceived, and between "You" and "your Self". Typically, in our experience, the perceiver and the perceived can be considered to be two distinct entities. Even when we look at ourselves, in a mirror for instance, we are really looking at a duplicate of ourselves. A seperate observer can see both the original and the duplicate, the reflection. In our model, the perceiver is not looking at a duplicate of his thoughts, but at the original. A separate observer would see that there is only one instance of the Self and its thoughts, the same instance that the perceiver is looking at.

Typically in issues of perception, the 'I' that perceives and the 'Self' that voices the comments are both on the inside, and the object being perceived is on the outside. Similarly a video camera and the television-monitor that it is attached to are typically thought of as the perceiving system, and the environment is the perceived. However, when the camera is pointed at the monitor, the monitor too becomes part of the environment. Similarly, with conscious thought, the thoughts themselves, and the Self that articulates them, become part of the environment for the perceiver, when he perceives them. The standard distinction between internal and external becomes untenable.

A long-standing conundrum in the field of Artificial Intelligence that this model resolves is the issue of "meta-level" reasoning, or reasoning about reasoning. The standard argumentruns: if we have reasoning by mental agents at one level being observed and reasoned about by agents at a higher level, who watches the reasoning of those higher agents - other agents at a still higher level? And is so, who watches their reasoning in turn? The issue of meta-level reasoning has always appeared to involve an infinite regress. Through our single-spokesman model, we can see that all conscious thought, even thoughts about thoughts, might be always at the same level, the level of the spokesman. This level, in turn, would be one level below all the commentators (keeping the convention that the observer is a level "above" the observed), some of whom are watching for patterns among the thoughts expressed by the spokesman. We can properly speak of these watchdog commentators as engaging in meta-level reasoning, but their thoughts are voiced at the same level as everyone else's, rather than at some privileged higher level. The faulty assumption which led to the paradox was that a reasoner and his reasoning would necessarily be at the same level - in fact, it appears that in the architecture of the conscious, the reasoner is always a level above his own reasoning.

Another puzzle that seems to be solved by our model is that of the occasional subjective experience, mentioned in the math exam case study, that some sort of Muse, separate from ourselves, is providing the answers.

In terms of our model, we can speculate that this feeling arises exactly when the thinking is done by a single commentator, and first comes into conscious awareness in a finished form. By contrast, when the answer is the cooperative effort of several commentators, using the spokesman as an interface during the intermediate steps, one should get the subjective experience that he himself, not some independent Muse, figured out the answer. Similarly, in human organizations, if one member of a committee submits a finished report before discussion has begun, it would seem odd to attribute the report to the committee.

## OPEN QUESTIONS AND SPECULATIONS

As always in studying nature, answering one question raises several new ones. In our case, having articulated our model, we find that on certain aspects our hypothesis is incomplete. We need, for instance, to extend our model slightly to handle one issue left unresolved so far. When several commentators send notes simultaneously, how is it determined which the spokesman will read first? Without loss of generality we could assign this scheduling function to the spokesman, but as this scheduling is an important function in its own right, with many questions surrounding it, I think it will be more fruitful to speak of the scheduler as a separate entity.

We know very little about the scheduler's decision procedure. It seems to prefer follow-on thoughts from the same or related sources, giving the sensation of a "cascade," or "train of thought". One can wonder, however, whether the scheduler even does the prioritizing at all, or whether thoughts came from their sources with some kind of priority-level or urgency already attached. These are open research questions, although the literature on shift-of-attention might be relevant.

One aspect of conscious thought that an expanded model might consider is the interrelation between memory and consciousness. Subjectively it appears that only those thoughts which pass through consciousness are candidates for recall later. It would appear that there is some sort of transcript kept of the events that take place in consciousness, but we have no case studies available at this time to delineate the exact relationship.

A major direction for expansion of the model would be to include stimuli from the outside world. While the model we have proposed may be informative for conscious thought and the cycles it involves, the model makes no real attempt to analyze the channels by which the outside world is pipelined into our consciousness. It merely claims that however news of the outside world reaches our minds, when we are conscious of the news, it is because that news is announced in the same auditorium as our own thoughts, and to the same audience of potential commentators.

## SUMMARY AND CONCLUSIONS

In summary, we have arrived at a rather unprecedented architecture for the phenomenon of conscious thought, based on considerations raised by our case studies. The example of reading to ourselves focused attention on the voice in each of our heads, and brought out the point that on certain occasions our minds can become crowded with more thoughts than we can handle. The experience with peak performance during mathematics competitions provided an extended example of the contrasts between various modes of mental functioning, and illustrated the "public" nature of conscious mental thought. Finally, the incident of being caught in an infinite loop gave a rare glimpse into the pathologies that our mind is susceptible to, and constrained us to propose models that exhibit the structural features necessary to support such pathological functioning.

Our discussion suggests that our mental life is populated by numerous commentators, all of whom are exposed to the contents of our consciousness, and at times comment on it. When we are engaged in conscious thought, however, the contents of our consciousness are simply the prior comments made by individual commentators. Each remark, rather than being voiced by the commentator itself, is voiced by a single, shared spokesman - the voice in our heads - and thus becomes an object for further comment, by the other commentators and even by the original commentator itself.

With this model in mind, we can pose further research questions, such as the nature and number of the commentators, and the type of decision-procedure involved in prioritizing the various thoughts. We can ask about the relationship between a thought being voiced by the spokesman, and a thought becoming a part of the transcript of our mental life. And we can hope to leverage our understanding of the process of conscious thought into an understanding of the final stages involved in being conscious of the world outside. But even though the answers to these questions are still ahead, with the development of this model we have taken a step down the road toward resolving the paradoxes of our conscious mental life, and toward illuminating what it is that is going on when you hear your "Self" think.